Introduction to Big Data

Lesson 1.2 Computer science survival kit

Arthur Katossky & Rémi Pépin

Tuesday, March 29, 2023

Computer science survival kit

A computer can be abstracted by four key components:

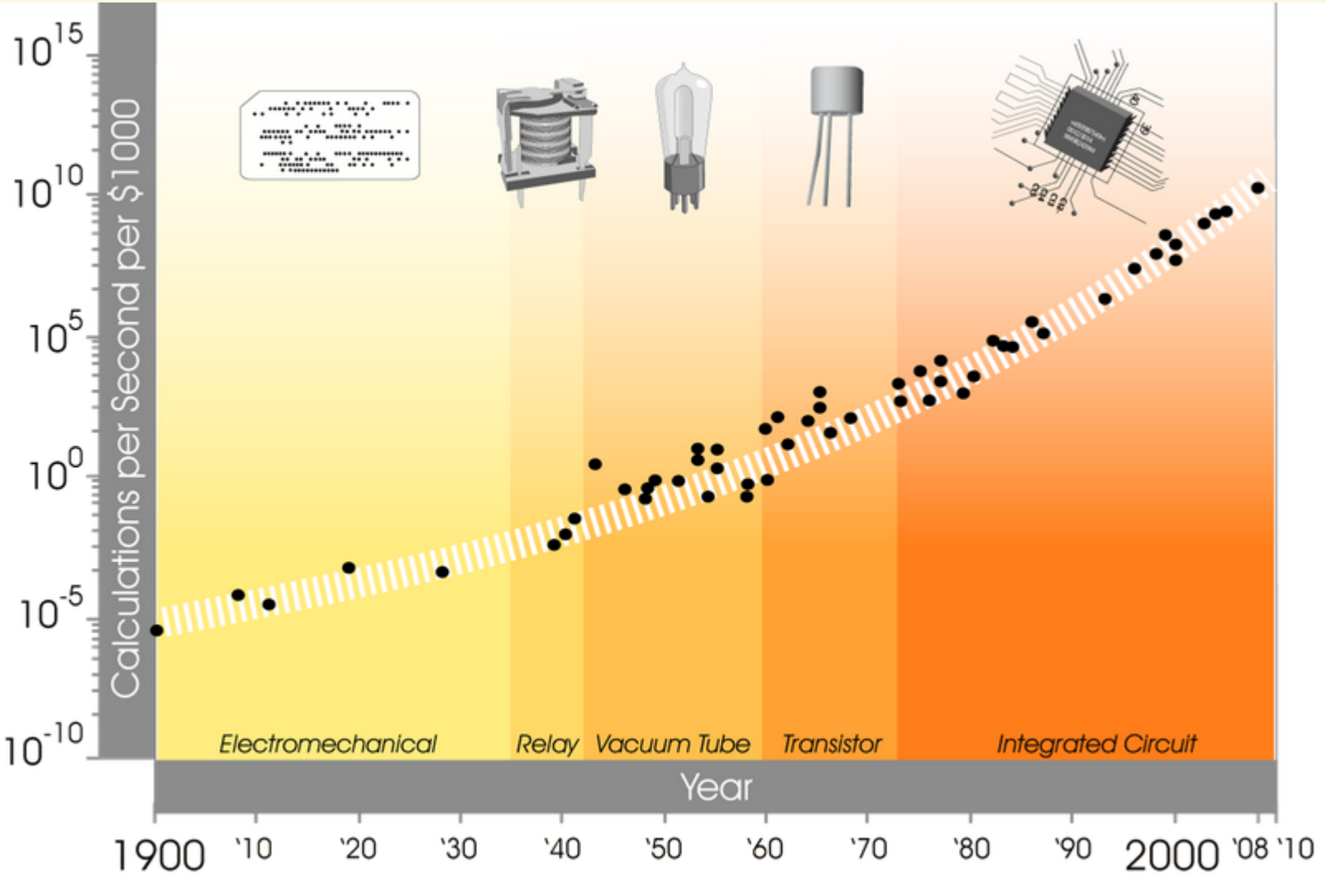
- processing (FR: capacité de calcul)
- memory (FR: mémoire vive)
- storage (FR: stockage)
- wiring / network (FR: réseau)

Processors

Processors

What is important for us to know?

- computation happens ***physically*** in **transistors**
- **programs must be converted to machine-code**, i.e. to a list of formatted instructions that the processor can execute
- **instructions are stored sequentially** in one or several stacks / threads / queues
- one program / application may access several treads (**multi-threading**)
- **processors can be specialized** (e.g. GPU)
- performance is measured in number **operations** or **instructions per second** (FLOPS, IPS)

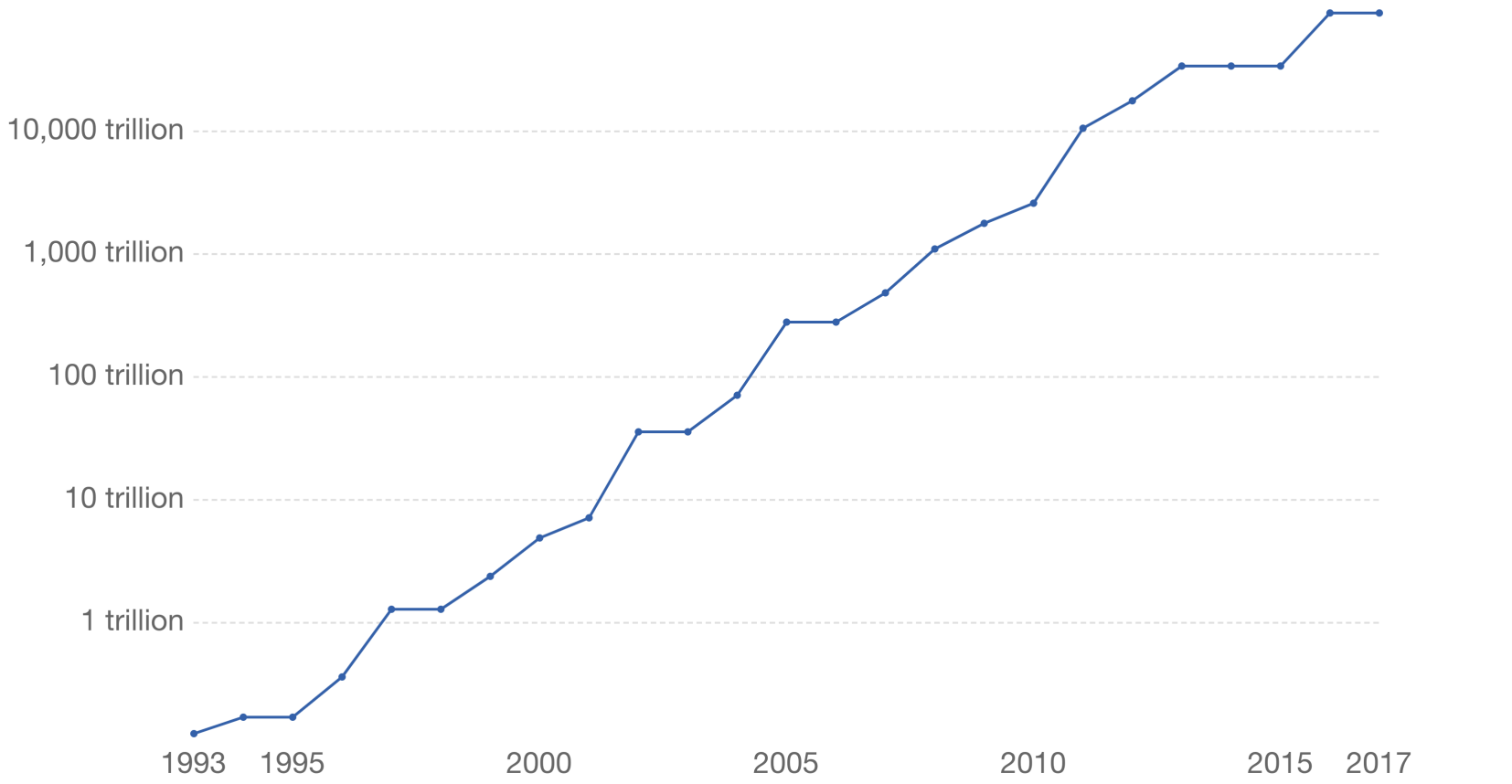


Source: Kurzweil (link) via Our World in Data (link)

Supercomputer Power (FLOPS)

The growth of supercomputer power, measured as the number of floating-point operations carried out per second (FLOPS) by the largest supercomputer in any given year. (FLOPS) is a measure of calculations per second for floating-point operations. Floating-point operations are needed for very large or very small real numbers, or computations that require a large dynamic range. It is therefore a more accurate measured than simply instructions per second.

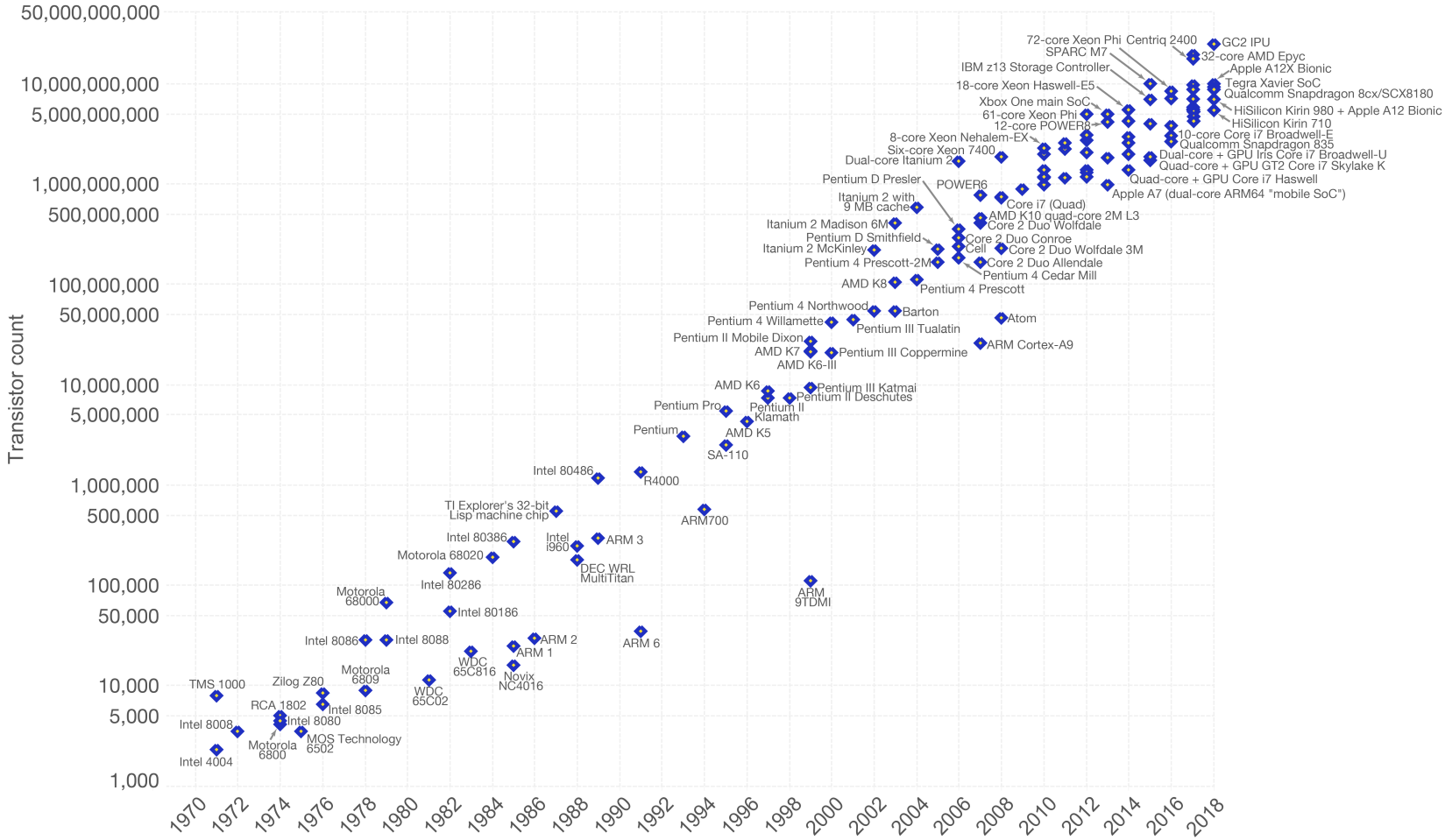
Our World in Data



Source: TOP500 Supercomputer Database

Moore’s Law – The number of transistors on integrated circuit chips (1971-2018)

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are linked to Moore's law.



Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)
The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.

Processors

Processing is a limitting factor.

- processors are **the most expensive part** of the hardware at buy time
- energy consumption of the processors are the **main cost of computation**
- one core can **only** perform **so many operations** per second
- it is **non-trivial** to **coordinate multiple processors** on the same complex task

Memory

Memory

What is important for us to know ?

- moving data around takes time
- memory is **fast** (ns)
- memory is **volatile** (lost in case of power interruption)
- memory-processor units are heavily **optimized**
- processors always access the the hard disk through **memory caching**

Source : https://aiimpacts.org/trends-in-dram-price-per-gigabyte/

Memory

Memory is a limitting factor.

- it is **non trivial** to work with data that **can't fit into memory**
- memory is **the second-most expensive part** of a computer at buy time
- memory is **shared** with other programs on the computer [1]

[1] Or with other users in case of shared server.


00111G

Storage

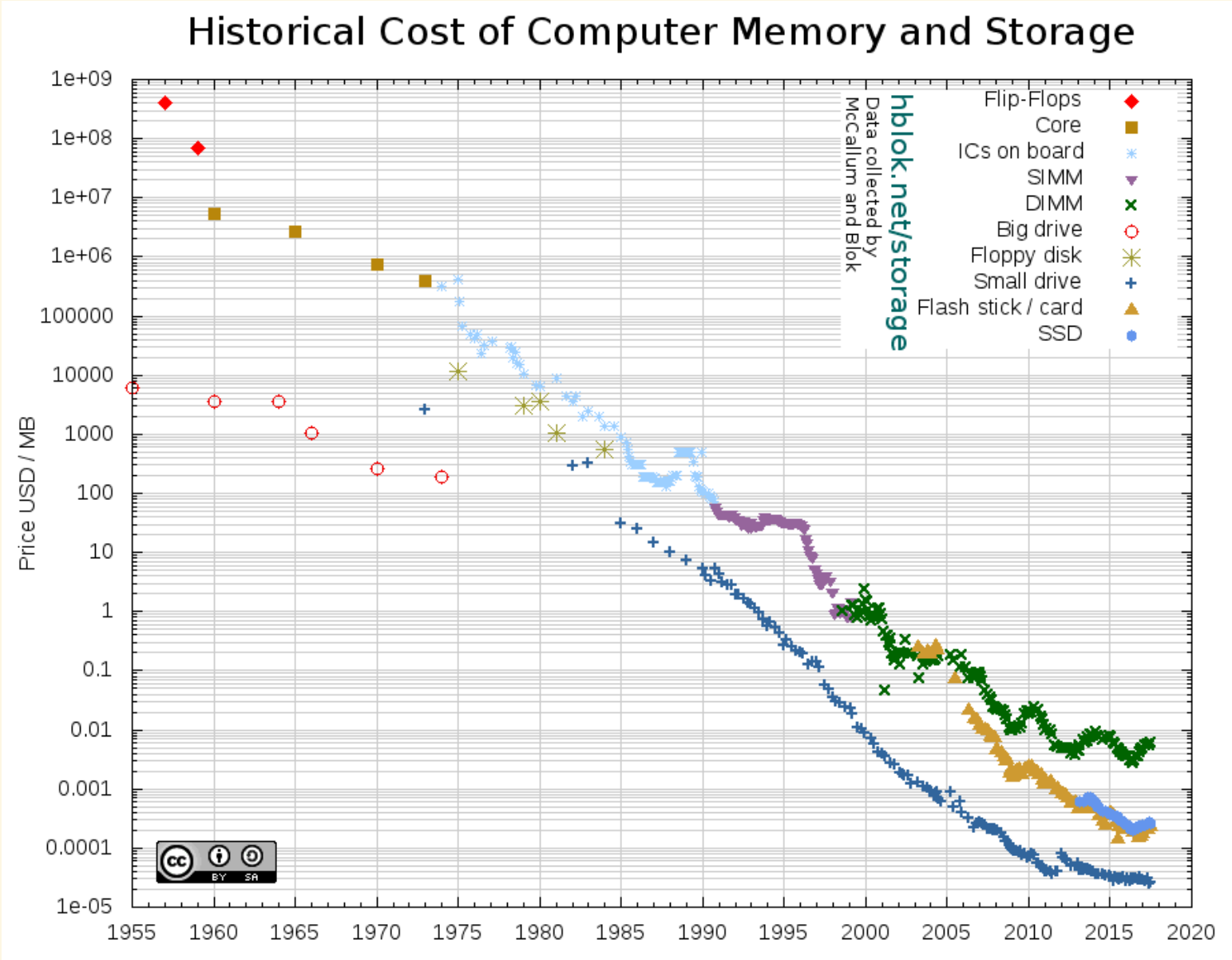
- for long-term storage of information
- can have **multiple forms**: transistor plates (USB sticks, SSD), magnetic disks (hard disk, floppy disk), physical engraving (vinyls, CDs, DVDs), magnetic tape, paper (books, punch cards, bar codes, QR codes), biological (DNA), etc.
- commonly referred to as "(disk / storage) space" or "(hard) disk" (even when no disk is involved)
- **non-volatile**, contrary to memory
- valuable properties:
  - **size**
  - **integrity**, resistance to degradation
  - **speed**, in read and write

Storage

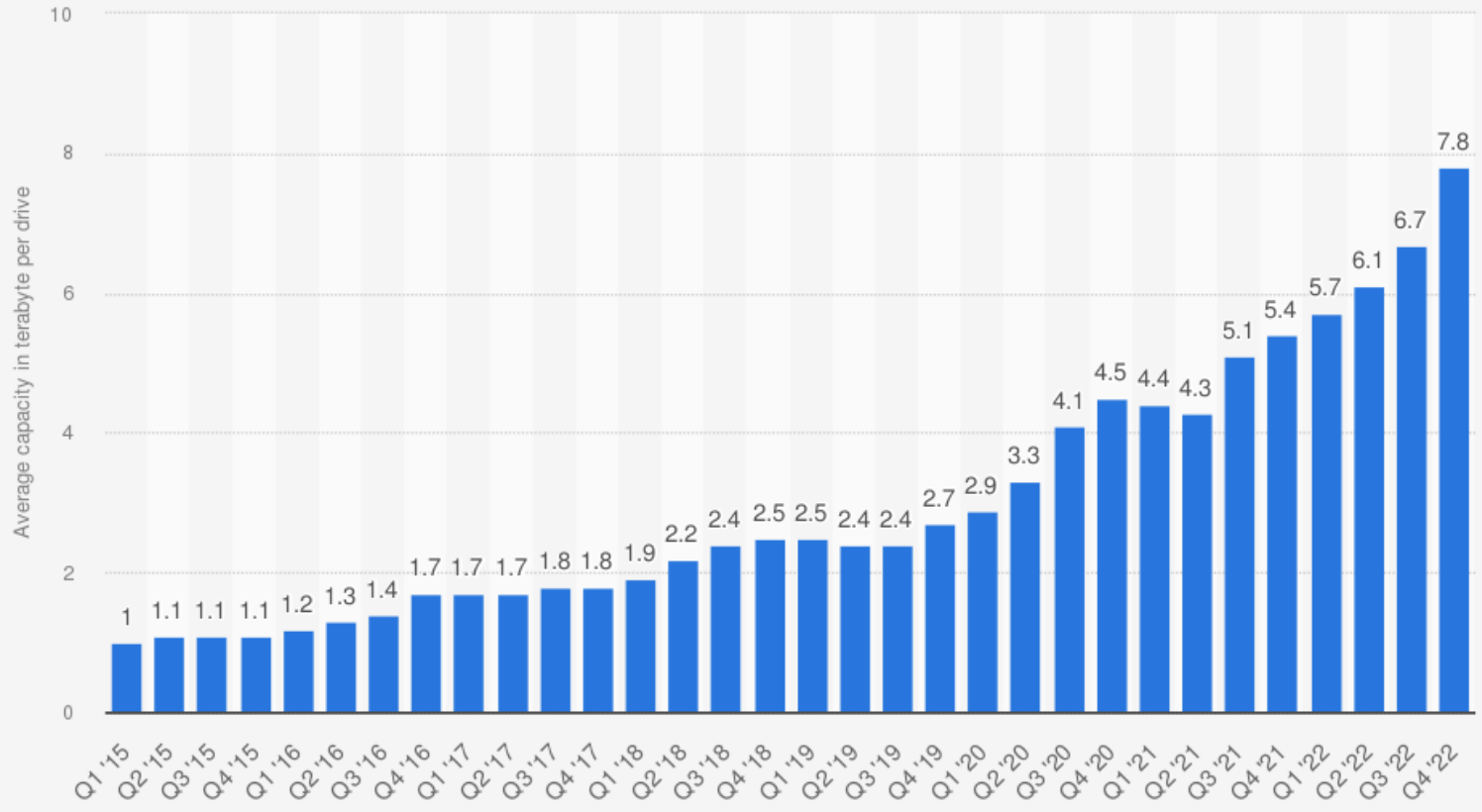
Storage is evaluated in **bytes** (B) or **octets** (o) and their mutiples:

Number of bytes	Symbol	Full name	Order of magnitude
10^3	ko	kilobyte	a vectorial icon
10^6	Mo	megabyte	a high resolution image, a book
10^9	Go	gigabyte	a high-resolution video
10^{12}	To	terabyte	the whole Friends series in high-resolution
10^{15}	Po	petabyte	Spotify database
10^{18}	Eo	exabyte	monthly Internet traffic

Since $2^{10} = 1024 \simeq 10^3$ and since storage is essentially binary, you often find the convention $1\mathrm{Mo} = 2^{10}\mathrm{ko}$. The official symbols Kio, MiB... never really caught on.


Historical Cost of Computer Memory and Storage
Price USD / MB
1e+09
1e+08
1e+07
1e+06
100000
10000
1000
100
10
1
0.1
0.01
0.001
0.0001
1e-05
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
Flip-Flops
Core
ICs on board
SIMM
DIMM
Big drive
Floppy disk
Small drive
Flash stick / card
SSD
hblok.net/storage
Data collected by McCallum and Blok
BY SA

Seagate's average capacity of hard disk drives (HDDs) worldwide from FY2015 to FY2022, by quarter (in terabyte per drive)



Source
Seagate

Additional Information:
Worldwide; Seagate; 2015 to 2022

Storage

What is important for us to know ?

- storage is **faillible**
- writing and reading is **slow** (ms)
- **data is always cached** (copied / charged into memory) for computation

Storage

Storage is a limiting factor.

- on a given computer, **you can only store so much**
- dealing with files **distributed** over multiple computers is **non-trivial**

Cost and speed are usually not an issue.

Network

- information transfer is **time-consuming**
- usually not an issue on a personal computer:
  - processors and memory are **closely integrated** in the same circuits
  - physical connection between memory and disk is **short** and **fast**
- becomes an issue with **remote** (or **distributed**) storage (or computing)

Network

For transferring volumes above 1 Po (1024 To) to their servers, Amazon actually sends a truck, which is faster than a fast Internet connection (https://aws.amazon.com/fr/snowmobile).

IN THE NEXT SECTION

Each of the basic components of a computer can cause issues.

- disk space is a limiting factor -> consider **distributed storage**
- even when data is relatively small, computation can be a challenge
- memory size and number of processors are a limiting factor -> consider **distributed computing**

But distribution is **non-trivial** (future lesson) and often not needed. Always consider first:

- consider simple **good practices** (next section)
- consider commercial **remote**, **cloud-based services** (future lesson)